

基于类心距离的模糊支持向量数据描述

王敏光 王 喆

(华东理工大学信息科学与工程学院 上海 200237)

摘 要 针对传统的支持向量数据描述模型忽略了样本分布的重要性,提出了基于类心距离的模糊支持向量数据描述算法,并将其应用在 UCI 机器学习数据库的二分类和多分类数据集中。该算法利用样本到两类中心距离的比值赋予样本权重,增大贡献度大的样本的权重,降低贡献度小的样本的权重,突出样本之间的差异性,从而提高了算法的分类效果。实验表明,该算法具有比传统支持向量数据描述更好的学习能力和分类效果。

关键词 模式识别,支持向量数据描述,权重

中图分类号 TP391

文献标识码 A

DOI 10.11896/j.issn.1002-137X.2016.5.042

Fuzzy Support Vector Data Description with Centers of Classes Distance

WANG Min-guang WANG Zhe

(School of Information Science & Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract Support vector data description(SVDD) ignores the importance of sample distribution. This paper proposed a new method called fuzzy SVDD with centers of classes distance, and it had been applied on the UCI data sets. The algorithm uses ratio of the distance of sample to the centers of two classes to give each sample a weight. The important samples' weights should be increased and the others should be not, which can highlight the difference of samples. The results show that our algorithm has better performance than SVDD.

Keywords Pattern recognition, Support vector data description, Weight

1 引言

处理数据分类问题时,负类数据缺失,两类分类器因为没有足够的负类数据进行训练会导致分类效果变差。而单类分类器恰恰相反,它只需要一类目标样本进行训练,描述出最优的目标类分布,因此被广泛运用于故障检测、医诊分析等领域,因为在这些领域中,负类样本的数据往往比较难收集,而正类样本却相对较多。支持向量数据描述(Support Vector Data Description, SVDD)是一种被广泛运用的单类分类器。SVDD 把原始空间的数据映射到高维空间,找到一系列支持向量,通过构建这些支持向量能够最大地包围正类样本,排除尽量多负类样本的超球面。针对 SVDD 的改进有很多,文献[1]利用样本局部密度信息为每一个样本赋予一个权重,实验结果表明,此算法的分类效果优于传统 SVDD。文献[2]同样利用样本的密度信息为样本加权,证明了样本分布对于分类边界确立的重要性。在文献[3]中,SVDD 与深度学习相结合,能够处理多类问题,并且没有严重的过拟合问题。为了克服 SVDD 忽略样本分布的问题,文献[4]通过构建两类的超球面来提高分类的准确率。文献[5]通过最近邻、帕森窗算法来估计样本的相关性密度,为样本赋予权重。文献[6]提出了一个基于 SVDD 的异常值边界检测方法,尝试不使用核技术,在原始空间中用样本到最近邻边界点的距离来构造一个

更为紧凑的决策超球面。文献[7]提出了基于改进 PCM 的模糊支持向量数据描述。针对大数据集分类时 SVDD 的时间复杂度太高的问题,文献[8]提出了快速 SVDD 算法,即利用边界检测的方法找到边界的样本点进行保留,删除冗余的类内样本点,从而降低时间复杂度。由于单核 SVDD 存在核函数、核参数选择的问题,文献[9]提出了多核 SVDD 算法。SVDD 已经被运用于许多领域。首先,SVDD 主要被广泛应用于检测区别正常样本的异常值样本,其中主要包括脸部识别、信用卡信用等级检测、噪声识别等等。文献[10]针对传统噪声源识别方法无法识别突变噪声源的不足,将 SVDD 扩展到多类分类来识别多类噪声源。其次,SVDD 也被大量应用于一些特殊的分类问题,比如不平衡分类问题。不平衡分类问题主要是类间样本不平衡问题,一类样本点较多,其他类样本点较少。医学诊断就是典型的不平衡问题,传统的两类分类器在解决不平衡问题中表现不优,而 SVDD 被证明适合处理不平衡问题。然而,SVDD 也有几个缺点,比如随着训练样本规模的增大,时间复杂度也会增大,不适合处理大规模数据集;对异常值点较为敏感;另外,在训练过程中 SVDD 默认所有的训练样本对于分类超球面的贡献度是一样,这是不合理的。我们知道,越是靠近类中心的样本点对于超球面的构建贡献度越小,而靠近边界的样本点对于超球面的构建贡献度较大。针对这个问题,本文增加边界样本点的权重,减小类内样本点的权重。

到稿日期:2015-05-07 返修日期:2015-08-02 本文受国家自然科学基金面上项目(61272198),上海市教育委员会科研创新项目(14ZZ054),中央高校基本科研业务费专项资金资助。

王敏光(1990—),男,硕士生,主要研究方向为机器学习、模式识别,E-mail:843178723@qq.com;王 喆(1981—),男,副教授,博士生导师,主要研究方向为机器学习、模式识别,E-mail:wangzhe@ecust.edu.cn。

基于以上分析,本文提出了基于类心距离的模糊支持向量描述(Sample-Center Fuzzy Support Vector Data Description, SCF-SVDD)。通过样本与类内中心、类间中心距离的比较,给出一种新的样本权重的计算方法。

2 SVDD 简要描述

大多数的 SVDD 改进算法都是基于单类的 SVDD 算法改进的,基算法只有一类样本参与训练,并没有其他类样本参与训练过程。考虑到本文的方法需要比较样本到不同类中心的距离,本文采用的是负类的 SVDD 模型(Negative Support Vector Data Description, N-SVDD)^[11]。假设一个数据集有两类样本 $\{x_1, x_2, \dots, x_i, x_{i+1}, \dots, x_{i+l}\}$,其中前 i 个样本为正类样本,后 l 个样本为负类样本。它的准则函数如下:

$$\begin{aligned} \min R^2 + C_1 \sum_i \xi_i + C_2 \sum_l \xi_l \\ \text{s. t. } \|x_i - \alpha\|^2 \leq R^2 + \xi_i, \|x_l - \alpha\|^2 > R^2 - \xi_l \\ \xi_i \geq 0, \xi_l \geq 0, \forall i, l \end{aligned} \quad (1)$$

其中, x_i 代表的是正类样本, x_l 代表的是负类样本。为了让模型更具鲁棒性,样本 x 到中心的距离不必严格地小于 R , 引入了参数 ξ_i, ξ_l 。 R 是超球面的半径。参数 C_1, C_2 用来平衡超球体积与误差。尽管 SVDD 对数据集提供了一个灵活、有弹性的边界描述,但是它也存在一些缺点。在一个数据集中,我们知道不同的样本点对于边界的贡献度并不相同,离样本中心越近的样本点对于边界的确立贡献度较小,而离边界越近的样本点对于边界的确立贡献度较大。然而 SVDD 默认所有的样本对于边界的确立贡献度相同,并没有考虑到样本分布的因素。因此,为了找到一个对目标类更为精确的描述,提出了一个改进的模糊 SVDD 算法,即通过引入一个新的权重参数来体现样本在数据集中的位置关系,增加边界样本点的权重,降低类内样本点的权重。

3 基于类心距离的模糊 SVDD

3.1 基于样本到类心距离的加权方法

上面提到,传统的 SVDD 模型并没有考虑到不同样本点对边界确立的贡献度问题。为了克服这个缺点,通过引入一个新的权重,提出了基于样本到类心距离的改进 SVDD。在介绍这个权重之前,首先定义了一个分类能力^[12]的概念:

$$\theta_{ip} = \frac{\|x_p^i - m_i\|_2}{\|x_p^i - m_i\|_2} \quad (2)$$

其中, m_i 表示的是第 i 类的样本均值, x_p^i 表示的是第 i 类第 p 个样本点。若 x_p^i 离本类样本均值(可以当作是本类中心点)越近,离其他类中心点越远,那么 θ_{ip} 值就越大。我们认为这样的样本点的分类能力高,也就是说容易被分类正确,因此这个样本点对于边界的确立贡献度不高,理应被分配一个较小的权重。当 θ_{ip} 较小时,代表这个样本距离边界较近,那么这个样本点对于边界的确立贡献度相对较好,权重需要增大。如图 1 所示,三角形代表正类样本,圆代表负类样本,五角星代表正类中心样本点,五边形代表负类中心样本点,黑色实心三角形代表正类边界样本点,黑色实心圆代表负类边界样本点。根据式(2),可以清楚地得到类内样本点的分类能力要大于边界样本点的分类能力,因此需要降低类内样本点的权重,增大边界样本点的权重。权重公式如下:

$$\mu_{ip} = \exp\left(\frac{1 - \theta_{ip}}{1 + \theta_{ip}} / \sigma\right) \quad (3)$$

其中, σ 用来控制权重的变化速度。通过这个权重公式,可以清楚地知道两类之间处于边界处的样本的权重会被增大,而在类内中心点附近的样本点权重将被减小。

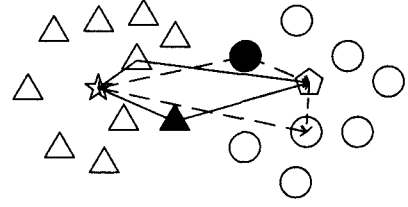


图 1 样本分类能力的效果图

3.2 基于类心距离的模糊 SVDD

经过 3.1 节就可以得到每一个训练样本的权重,把这个权重引入到 SVDD 模型中,新的准则函数如下:

$$\begin{aligned} \min R^2 + C_1 \sum_i \mu_i \xi_i + C_2 \sum_l \mu_l \xi_l \\ \text{s. t. } \|x_i - \alpha\|^2 \leq R^2 + \xi_i, \|x_l - \alpha\|^2 \leq R^2 - \xi_l \\ \xi_i \geq 0, \xi_l \geq 0, \forall i, l \end{aligned} \quad (4)$$

参数 C_1, C_2 用来平衡超球体积与误差, μ_i, μ_l 为样本权重。与传统准则函数(1)不同,式(4)中用权重 μ_i, μ_l 分别乘上松弛因子 ξ_i, ξ_l , 这样就为每一个样本加上了一个权重系数。那些在边界点附近的样本点被赋予一个大的权重,因此,为了最小化准则函数(4),超球面的构建重心会转向那些边界的样本点。另一方面,随着类内样本点权重的减小,这些样本点对边界的影响将会降低。为了解决这个最优化问题,引入了拉格朗日函数:

$$\begin{aligned} L(R, a, \xi, \alpha, \gamma) = R^2 + C_1 \sum_i \mu_i \xi_i + C_2 \sum_l \mu_l \xi_l - \sum_i \alpha_i \{R^2 + \xi_i - \|x_i - a\|^2\} - \sum_l \alpha_l \{ \|x_l - a\|^2 - R^2 + \xi_l \} - \sum_i \gamma_i \xi_i - \sum_l \gamma_l \xi_l \end{aligned} \quad (5)$$

求其偏导,新的公式和约束如下:

$$\begin{aligned} L = \sum_i \alpha_i (x_i \cdot x_i) - \sum_l \alpha_l (x_l \cdot x_l) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) + \\ 2 \sum_{i,j} \alpha_i \alpha_j (x_l \cdot x_j) - \sum_{l,m} \alpha_l \alpha_m (x_l \cdot x_m) \\ a = \sum_i \alpha_i x_i - \sum_l \alpha_l x_l \\ \text{s. t. } 0 \leq \alpha_i \leq \mu_i C_1, 0 \leq \alpha_l \leq \mu_l C_2 \end{aligned} \quad (6)$$

其中, α 为拉格朗日因子。在式(6)中对输入空间的样本点进行核映射处理,使其投射到高维的核空间,得到 $(\phi(x_i) \cdot \phi(x_j)) = (x_i \cdot x_j)$, 其中 ϕ 代表映射函数;最后用核函数 $K(x_i \cdot x_j)$ 来表示 $(\phi(x_i) \cdot \phi(x_j))$ 。已有的核函数有很多,常用的 3 种包括线性核、多项式核和高斯核函数。在本文实验中采用了高斯核,它的形式如下:

$$K(x_i \cdot x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \quad (7)$$

高斯核函数不依赖于样本的位置,它只利用了样本之间的距离关系。系数 σ 是个宽度参数,控制了函数的径向作用范围。通过式(6)中的约束条件,可以清楚地看到传统 SVDD 和模糊 SVDD 的区别。传统的 SVDD 中 α_i, α_l 的上界是固定的 C_1, C_2 值,然而模糊 SVDD 中 α_i, α_l 的上界是动态变化的,根据每一个样本不同的权重,它们对应的上界分别是 $\mu_i C_1, \mu_l C_2$ 。我们知道,SVDD 边界是通过拉格朗日因子线性组合得到的,这样通过确立不同样本的拉格朗日因子的动态上界就能够体现不同样本的差异性,能够区分不同样本对于模型的贡献度。

4 实验

4.1 UCI 数据集

在介绍实验策略和实验效果前,首先介绍本文用到的数据集。UCI 数据集是常用的标准测试数据集。本节列出本文实验中用到的 UCI 机器学习数据库中的二分类和多分类数据集。下面的实验结果都是通过在 UCI 上测试得到的,具体数据集和其相关信息如表 1 所列。

表 1 UCI 数据集

数据集	样本总数	类别数	每类样本个数
Iris	150	3	50/50/50
Wine	178	3	59/71/48
Glass	214	6	70/76/17/13/9/29
Heart	270	2	150/120
Balance	625	3	288/288/49
Water	116	2	65/51
Ionosphere	351	2	225/126
Sonar	208	2	97/111

4.2 实验策略设置

为了验证模糊 SVDD 算法在分类中的效果,下文采用 5 轮交叉验证的方法将其与传统 SVDD 模型、N-SVDD 和 D-SVDD^[1]在 UCI 机器学习数据库中的二分类和多分类数据集的分类效果进行比较。实验中使用的核函数是高斯核函数,

其中核参数采用文献[13]的设定方法,即 $\sigma^2 = \frac{1}{n^2} \sum_{i,j=1}^n \|x_i - x_j\|^2$, n 为训练样本个数。 $C_1 = C_2 = \frac{1}{\epsilon_n}$, 其中 $\epsilon = [10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3]$ 。

实验中用来评价分类算法的指标为正类准确率(True Positive Rate, TPR)、负类准确率(True Negative Rate, TNR)、g-means。其中:

$$TPR = \frac{\text{正类中被正确分类的样本个数}}{\text{正类样本个数}}$$

$$TNR = \frac{\text{负类中被正确分类的样本个数}}{\text{负类样本个数}}$$

$$g\text{-means} = \sqrt{TPR \cdot TNR}$$

4.3 实验结果

本节将本文的算法和 3 个对比算法在相同的核参数设置和相同 C 参数取值范围内分别进行 5 轮交叉验证实验。

4 个算法在 UCI 数据集上的 TPR、TNR 效果具体如表 2 所列,可以看到本文的算法 SCF-SVDD 在大部分数据集上的分类效果比 3 个对比算法出色,有个别数据集效果并不是最优,但与最优的结果接近,效果相当;在目标类的分类效果表现较优,在负类上的分类效果虽然并不占优,但基本也是在前列。总体来说,SCF-SVDD 具有更好的分类效果。

表 2 算法在 UCI 数据集上的 TPR、TNR

数据集	SVDD		N-SVDD		D-SVDD		SCF-SVDD	
	TPR(%)	TNR(%)	TPR(%)	TNR(%)	TPR(%)	TNR(%)	TPR(%)	TNR(%)
Iris(1)	90.00	100.00	96.00	93.87	82.00	<u>100.00</u>	97.60	99.20
	(0.10)	(0.00)	(5.06)	(7.76)	(0.13)	(0.00)	(3.20)	(1.60)
Iris(2)	84.00	95.00	88.00	<u>95.47</u>	84.00	95.00	94.40	93.33
	(0.08)	(0.00)	(9.12)	(4.10)	(0.11)	(0.00)	(5.99)	(2.38)
Iris(3)	74.00	99.00	78.40	96.27	74.00	<u>99.00</u>	84.00	95.73
	(0.11)	(0.00)	(8.98)	(2.29)	(0.11)	(0.00)	(7.59)	(5.93)
Wine(1)	72.67	89.58	76.67	93.55	76.67	89.24	79.33	<u>94.44</u>
	(4.90)	(4.53)	(11.93)	(2.76)	(10.87)	(0.70)	(9.75)	(1.86)
Wine(2)	82.79	56.64	75.00	85.28	<u>89.33</u>	46.17	72.22	<u>90.55</u>
	(7.12)	(15.01)	(5.55)	(7.74)	(5.96)	(11.38)	(4.65)	(6.17)
Wine(3)	70.00	89.85	69.17	<u>96.04</u>	<u>88.00</u>	90.77	84.17	91.70
	(3.12)	(6.35)	(1.55)	(4.60)	(8.37)	(4.14)	(6.12)	(4.97)
Glass(1)	90.29	100.00	97.14	95.23	88.57	<u>100.00</u>	100.00	99.27
	(4.64)	(0.00)	(5.72)	(2.68)	(0.06)	(0.00)	(0.00)	(1.07)
Glass(2)	83.68	100.00	<u>99.47</u>	94.60	92.50	<u>100.00</u>	97.90	97.80
	(5.37)	(0.00)	(1.05)	(5.57)	(0.08)	(0.00)	(3.07)	(2.56)
Glass(3)	80.00	<u>99.69</u>	88.89	86.14	80.00	97.96	100.00	93.76
	(17.78)	(0.41)	(22.22)	(9.69)	(0.20)	(0.01)	(0.00)	(3.48)
Glass(4)	68.57	<u>98.61</u>	100.00	87.90	60.00	97.31	100.00	93.13
	(27.70)	(1.15)	(0.00)	(8.95)	(18.95)	(0.01)	(0.00)	(5.05)
Glass(5)	84.00	<u>99.80</u>	100.00	61.79	76.00	80.00	100.00	82.69
	(8.00)	(0.39)	(0.00)	(20.96)	(29.39)	(39.90)	(0.00)	(15.4)
Glass(6)	70.67	<u>99.68</u>	90.67	94.62	60.00	99.24	100.00	97.54
	(9.04)	(0.43)	(18.67)	(3.44)	(8.43)	(0.94)	(0.00)	(2.04)
Heart(1)	80.53	30.83	50.93	<u>73.79</u>	<u>81.87</u>	35.83	63.47	72.00
	(8.42)	(6.77)	(11.73)	(14.57)	(0.06)	(0.04)	(6.95)	(9.04)
Heart(2)	77.00	31.60	57.67	68.89	76.00	30.13	57.33	<u>76.89</u>
	(7.99)	(8.73)	(11.09)	(8.49)	(12.89)	(10.29)	(3.43)	(2.85)
Balance(1)	89.40	67.48	80.97	<u>97.30</u>	80.69	78.58	89.58	94.40
	(3.15)	(1.29)	(10.71)	(2.16)	(4.46)	(1.90)	(4.83)	(4.64)
Balance(2)	84.58	68.55	84.03	<u>97.00</u>	82.07	79.58	90.69	92.95
	(3.50)	(7.31)	(7.31)	(1.41)	(4.33)	(0.88)	(2.04)	(1.93)
Balance(3)	48.00	61.70	62.40	75.98	<u>84.00</u>	26.94	57.60	<u>78.81</u>
	(6.69)	(9.94)	(12.29)	(4.68)	(8.94)	(2.78)	(11.20)	(5.19)
Water(1)	78.79	63.53	73.34	89.47	80.00	82.35	84.85	<u>90.53</u>
	(12.27)	(24.57)	(9.07)	(7.44)	(12.87)	(1.96)	(5.07)	(6.98)
Water(2)	77.69	<u>90.15</u>	89.77	89.00	80.00	84.00	90.00	87.50
	(7.46)	(1.57)	(5.22)	(3.74)	(16.26)	(4.69)	(5.75)	(7.91)
Ionosphere(1)	87.61	88.73	85.31	<u>97.14</u>	84.44	87.14	92.57	94.29
	(3.02)	(0.92)	(3.74)	(3.50)	(4.16)	(1.81)	(2.48)	(5.35)

(续表)

数据集	SVDD		N-SVDD		D-SVDD		SCF-SVDD	
	TPR(%)	TNR(%)	TPR(%)	TNR(%)	TPR(%)	TNR(%)	TPR(%)	TNR(%)
Ionosphere(2)	66.03	2.76	31.11	77.78	54.62	12.18	23.49	85.31
	(5.37)	(1.58)	(6.78)	(5.54)	(6.32)	(1.07)	(3.24)	(4.77)
Sonar(1)	65.72	49.55	44.90	78.20	59.00	60.18	51.43	78.21
	(7.12)	(10.14)	(8.27)	(1.79)	(10.84)	(2.33)	(10.11)	(10.10)
Sonar(2)	64.64	53.40	62.50	74.24	65.22	69.69	62.50	76.27
	(6.53)	(7.30)	(8.53)	(5.40)	(18.70)	(5.63)	(11.46)	(6.16)

注:其中黑色粗体为 TPR 值最优,下划线黑色粗体为 TNR 值最优。

SCF-SVDD 及 3 个对比算法在 UCI 数据集上的 g-means 表现如表 3 所列。可以看到,与表 2 稍有不同的是,在 g-means 的表现上,SCF-SVDD 的效果基本上是优于其他 3 个对比算法的,说明了本文的算法在两类分类问题中表现较优,证明了本文方法的可行性。

表 3 算法在 UCI 数据集上的 g-means

数据集	SVDD	N-SVDD	D-SVDD	SCF-SVDD
	g-means	g-means	g-means	g-means
Iris(1)	94.75	94.74	90.30	98.38
	(0.05)	(3.21)	(0.07)	(1.52)
Iris(2)	89.22	91.42	89.16	93.78
	(0.04)	(3.32)	(0.06)	(2.17)
Iris(3)	85.39	86.68	85.39	89.44
	(0.06)	(4.35)	(0.06)	(2.49)
Wine(1)	80.58	84.30	82.50	86.36
	(2.57)	(5.93)	(5.65)	(4.88)
Wine(2)	67.34	79.89	63.61	80.82
	(7.67)	(5.65)	(8.11)	(4.53)
Wine(3)	79.22	80.76	89.20	87.69
	(2.87)	(8.75)	(3.81)	(2.18)
Glass(1)	94.99	96.09	94.06	99.63
	(2.45)	(1.87)	(0.03)	(0.54)
Glass(2)	91.43	96.95	96.10	97.81
	(2.99)	(2.70)	(0.04)	(1.33)
Glass(3)	88.73	85.99	87.88	96.81
	(9.98)	(10.37)	(0.12)	(1.80)
Glass(4)	80.15	93.63	75.24	96.47
	(17.54)	(4.86)	(12.56)	(2.64)
Glass(5)	91.47	77.35	64.62	90.52
	(4.27)	(14.03)	(37.47)	(8.64)
Glass(6)	83.77	91.86	76.96	98.76
	(5.36)	(9.48)	(4.99)	(1.04)
Heart(1)	49.20	59.89	53.94	67.13
	(2.90)	(3.20)	(0.02)	(1.84)
Heart(2)	48.59	62.21	46.30	66.35
	(4.81)	(3.03)	(0.05)	(2.05)
Balance(1)	77.68	88.51	79.60	91.87
	(1.81)	(5.45)	(1.86)	(2.53)
Balance(2)	75.97	90.17	80.80	91.81
	(3.39)	(3.53)	(2.33)	(1.30)
Balance(3)	53.85	68.54	47.37	66.89
	(1.96)	(8.29)	(0.78)	(6.02)
Water(1)	68.72	80.82	80.92	87.47
	(14.07)	(6.26)	(5.68)	(2.70)
Water(2)	83.58	89.80	81.43	88.63
	(3.97)	(2.63)	(7.15)	(5.24)
Ionosphere (1)	88.15	90.97	85.80	93.39
	(1.27)	(1.23)	(2.55)	(3.19)
Ionosphere (2)	12.91	43.74	25.70	44.66
	(3.04)	(3.97)	(1.18)	(3.55)
Sonar(1)	56.39	58.99	59.30	62.60
	(3.30)	(5.40)	(5.08)	(4.39)
Sonar(2)	58.39	67.76	66.50	68.52
	(2.85)	(2.89)	(7.63)	(6.06)

注:黑色粗体为 g-means 最优。

统 SVDD 有一个较好的提升。从 g-means 表中可以发现,Iris、Heart、Ionosphere 和 Sonar 每一类当训练样本的 g-means 效果都优于其他 3 个对比算法。Wine、Glass、Balance、Water 的大部分 g-means 效果优于其他算法,个别类别当训练样本时,其 g-means 没有达到最优,但与最优的算法效果也是相当的。总体上,本文算法在 UCI 上具有一定的适用性。传统 SVDD 没有考虑到样本分布的问题,认为所有样本对于分类边界的贡献度相同,致使其对噪声点相当敏感。噪声点的存在对于分类边界的影响较大,会降低分类准确性。本文提出的模糊 SVDD 算法主要关注边界样本点,之所以会提升分类效果主要有下面两个原因:其一,通过拉格朗日因子我们知道,对 SVDD 分类边界有影响的是支持向量,而绝大多数的边界样本点都是支持向量,增大边界样本点的权重,提升了支持向量对分类边界的贡献度;其二,降低类内样本点的权重,也降低了类内噪声点对分类边界的影响。

结束语 本文提出了一个改进的 SVDD 算法即基于类心距离的模糊 SVDD 算法。这个算法赋予每一个训练样本一个权重,用来区分边界样本点和类内样本点的重要性。我们认为边界的样本点对于决策超球面的贡献度较大,相应的权重就较大;类内样本点对于决策超球面的贡献度较小,相应的权重就较小。UCI 数据集的实验结果表明,所提算法优于传统的 SVDD 模型,证明了边界的样本点在 SVDD 训练过程中对边界的确立影响很大,传统的 SVDD 对所有训练样本一视同仁,这显然是不可取的。

未来我们可以从 SVDD 在处理大规模数据中时间复杂度过高的问题出发,利用本文的方法,进行数据集的预处理,区分样本的重要性,进行训练样本的删减,保留边界的样本点,删除冗余的类内样本点,在基本不影响分类效果的情况下,减小时间复杂度,提高效率。

参考文献

- [1] Lee K, Kim D, Lee D, et al. Improving support vector data description using local density degree[J]. Pattern Recognition, 2005,38(10):1768-1771
- [2] Cha M, Kim J S, Baek J-G. Density weighted support vector data description[J]. Expert Systems with Applications, 2014,41(7): 3343-3350
- [3] Kim S, Choi Y, Lee M. Deep Learning with Support Vector Data Description[J]. Neurocomputing, 2015,165:111-117
- [4] Nguyen P, Tran D. Repulsive-SVDD Classification[C]// 19th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2015
- [5] K L, DW K, KH L, et al. Density-Induced Support Vector Data Description[J]. IEEE Transactions on Neural Networks, 2007, 18(1):284-289

在 TPR、TNR 或者是 g-means 上,本文算法的表现比传

(下转第 242 页)

结束语 支持向量机是计算密集型学习算法,不论 SVM 还是 Cascade SVM 算法,它们对大数据集的模型训练代价都较高,时间长。通过对 Cascade SVM 的并行策略进行深入的分析,结合 Spark 计算平台,实现了一种基于 Spark 的并行 SVM 训练算法,这个模型以数据划分和级联训练为基础,对 Cascade SVM 的训练结构进行了相应的改进,进而解决了层叠模型训练效率低的缺点。通过大量实验分析表明,该算法在准确率没有明显降低的前提下,能够提高分类速度,是 SVM 算法求解大规模数据集的一种有效解决方案。但是该算法还存在很多问题需要研究,例如当数据集的支持向量太多时算法训练效率不高,数据分块过多时准确率会有所下降,并行设计的合理性没有严谨的数学证明和理论推导。

在接下来的学习中,将对算法的理论支持和训练结构的优化做更加深入的研究,同时也会着重探讨 Spark 平台的参数调优,以期获得更好的训练效果,满足广大学者对采用支持向量机训练大规模数据集的需求。

参考文献

- [1] Vapnik V N. The Nature of Statistical Learning Theory[M]. Springer New York, 1995: 988-999
- [2] Chang C C, Lin C J. LIBSVM: a Library for Support Vector Machines[J]. ACM Transactions on Intelligent Systems & Technology, 2006, 2(3): 389-396
- [3] Dong J X, Krzyzak A, Suen C Y. Fast SVM training algorithm with decomposition on very large data sets[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(4): 603-618
- [4] Lin C Y, Tsai C H, Lee C P, et al. Large-scale logistic regression and linear support vector machines using spark[C]//2014 IEEE International Conference on Big Data. IEEE, 2014: 519-528
- [5] Zhang Wei, Zhang Gong-xuan, Wang Yong-li, et al. Research on parallel SVM algorithm based on CUDA[J]. Computer Science, 2013, 40(4): 69-72 (in Chinese)
- 张巍, 张功萱, 王永利, 等. 基于 CUDA 的 SVM 算法并行化研究[J]. 计算机科学, 2013, 40(4): 69-72
- [6] Graf H P, Cosatto E, Bottou L, et al. Parallel Support Vector Machines: The Cascade SVM[C]//Advances in Neural Information Processing Systems(NIPS). 2004: 521-528
- [7] Sun Zhan-quan, Fox G. Study on Parallel SVM Based on MapReduce[C]//The 2012 International Conference on Parallel and Distributed Processing Techniques and Applications. Las Vegas NV USA, 2012
- [8] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[J]. Proceedings of Operating Systems Design and Implementation(OSDI), 2004, 51(1): 107-113
- [9] Zhang Peng-xiang, Liu Li-min, Ma Zhi-qiang. Research of parallel SVM algorithm based on MapReduce[J]. Computer Applications and Software, 2015, 32(3): 172-176 (in Chinese)
- 张鹏翔, 刘利民, 马志强. 基于 MapReduce 的层叠分组并行 SVM 算法研究[J]. 计算机应用与软件, 2015, 32(3): 172-176
- [10] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets[C]//Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing USENIX Association, 2010: 10
- [11] <http://spark.apache.org>
- [12] Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing[C]//Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. 2012: 141-146
- [13] Guo Xin-xin. SVM optimization algorithm based on Distributed Computing[D]. Xi'an: Xi'an Electronic and Science University, 2014 (in Chinese)
- 郭欣欣. 基于分布式计算的 SVM 算法优化[D]. 西安: 西安电子科技大学, 2014
- [14] <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>

(上接第 233 页)

- [6] Guo S M, Chen L C, Tsai J S H. A boundary method for outlier detection based on support v-vector domain description[J]. Pattern Recognition, 2009, 42(1): 77-83
- [7] Zhang Y, Chi Z, Li K. Fuzzy multi-class classifier based on support vector data description and improved PCM[J]. Expert Systems with Applications, 2009, 36(5): 8714-8718
- [8] Hu Chen-long, Zhou Bo, Hu Jing-lu. Fast support vector data description training using edge detection on large datasets[C]//2014 International Joint Conference on Neural Networks(IJCNN). IEEE, 2014: 2076-2182
- [9] Wu Ding-hai, Zhang Pei-lin, Wang Huai-guang, et al. One-class classification method based on multi-kernel support vector data description[J]. Computer Engineering, 2013(5): 165-168 (in Chinese)
- 吴定海, 张培林, 王怀光, 等. 基于多核支持向量数据描述的单类分类方法[J]. 计算机工程, 2013(5): 165-168
- [10] Gao Zhi-hua, Ben Ke-rong. Study on acoustic sources identification based on multi-svdd[J]. Computer Science, 2012, 39(11): 233-236 (in Chinese)
- 高志华, 贲可荣. 基于多分类支持向量数据描述的噪声源识别研究[J]. 计算机科学, 2012, 39(11): 233-236
- [11] Tax D M J, Duin R P W. Support Vector Data Description[J]. Machine Learning, 2004, 54(1): 45-66
- [12] Chai J, Liu H, Bao Z. Generalized re-weighting local sampling mean discriminant analysis[J]. Pattern Recognition, 2010, 43(10): 3422-3432
- [13] Tsang I W, Kocsor A, Kwok J T. Efficient kernel feature extraction for massive data sets[C]//Kdd Proceedings of ACM Sigkdd International Conference on Knowledge Discovery & Data. 2006